SRI International

Technical Note 526 • September 1993

BUILDING AND USING SCENE REPRESENTATION IN IMAGE UNDERSTANDING

Prepared by:

H. Harlyn Baker
Senior Computer Scientist

Artificial Intelligence Center
Computing and Engineering Sciences Division

333 Ravenswood Avenue • Menlo Park, CA 94025-3493 • (415) 326-6200 • FAX: (415) 326-5512 • Telex: 334486

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
SEP 1993		**00-09-1993 to 00-09-1993**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Building and Using Scene Representation in Image Understanding	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **SRI International,333 Ravenswood Avenue,Menlo Park,CA,94025**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**		**13**	

Building and Using Scene Representations in Image Understanding

H. Harlyn Baker*
Artificial Intelligence Center
SRI International
Menlo Park, CA 94025, USA

1. SUMMARY

The task of having computers able to understand their environments through direct imaging has proved to be formidable. With its beginnings about 30 years ago (1), the field of computer vision has grown as a major part of the pursuit for artificial intelligence. Most elements of this pursuit – language understanding, reasoning and planning, speech – are very difficult challenges, but vision, with its high dimensionality of space, time, scale, color, dynamics, and so forth, may be the most challenging. Early attempts to develop computer vision focused on restricted situations in which it was feasible to provide the computer with fairly complete descriptions of what it would encounter. In such cases, single images provided the sensory information for analysis. As the domains of application grew, the requirements for more competent descriptions of the world increased. Dealing with three-dimensional (3D) dynamic structures (the real world) from 3D dynamic platforms (we humans) calls for greater capabilities on both the analysis and synthesis sides of the issue. The analysis side is the processing of sensory data for such tasks as recognition and navigation, and a number of techniques are discussed here for dealing with these two-, three-, and higher-dimensional data. The synthesis side is the construction of 'internal' descriptions of what is seen in the environment – constructed now so that they may be used subsequently for the above tasks. This latter issue is the underlying theme we pose in this paper – developing representations from vision that will later enable effective automated operation in our 3D dynamic environments.

2. INTRODUCTION

Vision, which appears so easy for all of us, has proved to be an extremely complex task when addressed with computers. Despite early expectations in the field for realization of machine vision capabilities, it has grown to occupy a large proportion of the continuing artificial intelligence research effort. Understanding the coarse structure, let alone the nuances, of our environment continues to be a large and, in many parts, elusive challenge.

*The SRI research discussed here has been sponsored by DARPA under contracts DACA-76-85-C-0004, DACA-76-90-C-0021, and DACA-76-92-C-0003, and by Fujitsu System Integration Laboratory.

2.1 Knowledge for Analysis

A major component of the vision efforts seen today still parallels approaches taken throughout the years – the building in to the system of specific knowledge of the domain it will encounter. Vision does not take place without memory. As sighted individuals, we have a great deal of expertise, accumulated over years of observing and interacting with our 3D dynamic environments. Undoubtedly, certain capabilities appear with us at birth. Experience, however, and the memory that it accumulates, is equally critical to our performance. It enables us to rapidly and robustly interpret situations and events, recognize the familiar, and react opportunely to what we see. Since experience appears so necessary to our performance, it seems essential that a computer charged with seeing also have access to some equivalent sort of background knowledge. Although seldom enunciated, how this knowledge is given to the system, how it is represented, and how it is used in analysis of the visual imagery turn out to be principal issues in computer vision.

These knowledge issues occur at all levels of the analysis, from deciding what useful information from small parts of individual images to extract for subsequent processing (e.g., brightness values, gradients, contour elements), to considering what is relevant for identifying a striding distant silhouette as one's Uncle Bob. At some levels of the analysis there are generally accepted definitions of the knowledge that is appropriate (for example, the use of spatial-frequency-tuned filters), but, mostly, very little is understood and very little is agreed upon about these matters.

2.2 Representational Limitations

My discussion here relates to this knowledge-source issue. I phrase it as building and using computational representations in the task of understanding what is presented in an image of a scene. I present a number of pieces of work, indicating the capability they were designed to provide, the role of this capability in a vision system, and the level of initial-state knowledge provided to the system along with its ability to augment this through time. The main point I draw out is that all computer vision systems begin with an alphabet of operational primitives used to represent the image data. They have a vocabulary of combinations of these that they can deal with for scene interpretation. The capability of the system is set by its expressive power in this vocabulary, while its utility in a broader context is determined by the breadth of these definitions and its ability to grow beyond their limiting bounds. The

latter issue pushes up against generic 'learning,' an area of artificial intelligence probably unparalleled in both its potential and the ratio of its promise to its realization.[1] However, the issue of a system's repertoire of expression – its ability to build representations from imaged data and use them in understanding the visual situation – provides a key measure of its contributions: its contribution in solving the particular problem it addresses as well as its contribution to the computer vision task in general.

Two major determinants of the capabilities of a vision system are (1) the modes of imaging used, and (2) the elements on which it bases its analysis. In the next section I will provide a reference framework for these by discussing the principal modes of image data acquisition (single images, binocular stereo, and dynamic sequences) and the two choices for processing styles – homogeneous versus structured. The comparisons of image understanding systems I make in the following sections will be framed by these categories.

3. IMAGING MODALITIES

Imagery for scene analysis comes in three principal forms: monocular views; binocular views, and multi-image sequences of views – looking at a photograph, looking with your two eyes without being able to move your head, and the general situation of two eyes on a mobile head. Each form of data contributes differently to the scene representation and image understanding tasks.

3.1 Dynamic Scenes

Image sequences may provide information about scene dynamics (other moving objects), or give differing perspectives on a scene viewed as the sensor moves around. This is a mode of operation that people are clearly very capable of using, as we observe our dynamic world and move around in it, exploring. The relatively new area of 'active' vision (as in a sensor that adjusts its perspective to satisfy its requirements) studies acquiring and exploiting these sorts of data. Since, from the viewpoint of survival, anything that is in motion in our vicinity is of special interest to us, the analysis of dynamic imagery may be expected to play a critical part in a computer vision system.[2] Taking the more active role in data acquisition – moving around and collecting information from a variety of perspectives – leads to considerably more robust and more precise scene measurements. The cost is considerably more processing.

3.2 Binocular Viewing

What a single moving sensor does not provide is precise 3D measurement of moving objects. To determine the three-space position of an object requires seeing it from several (at least two) known perspectives simultaneously. A moving object viewed by a moving sensor is viewed from only one perspective at any instant.

Binocular views, image pairs captured simultaneously from different locations (as the eyes provide), can give sufficient information to enable 3D interpretation of both static and dynamic elements of a scene. That is, simple triangulation (back projection) can be applied to corresponding points in two images from known viewing positions to determine the location of the observed point in three-space. The biggest problem in stereo – one that has been with us from the beginning – is developing reliable techniques for determining which point in one image corresponds to a point in the other. This is the 'correspondence' problem – matching elements[3] between views. Although static binocular viewing is unusual – in human vision most binocular perception is dynamic – it is certainly effective, as viewing Figure 5 (subsection 6.3.3) will show. Depth is a powerful aid to scene understanding.

3.3 Single Images

With a stationary sensor viewing a nonchanging scene, a single snapshot view may be all that is available, and alone must be the basis for scene interpretation. That humans can operate with such a deficiency of information, for example in viewing photographs, lacking dynamics and explicit three-dimensionality, reveals the power of our processing and the value of memory and experience.

Most early theses in computer vision dealt with analysis of single images, and their failings immediately taught us the lesson of extensibility. Lacking access to the rich information of depth and motion, systems for single-image analysis were initialized with specific knowedge of the simple objects with which they could deal, and had no way to grow beyond this aside from reprogramming.

If all that is presented is a single image, and never in the context of a dynamic sequence, any interpretation will have to forego explicit temporal or 3D analysis. Since we presumably do not begin life with explicit knowledge of 3D structures, such as houses and cars, yet develop understanding of them over time (with both stereo and temporal data available), it is inconceivable that memory could operate without temporal analysis.

3.4 Processing Elements

A distinction within the different modes of operation that will be contrasted throughout this article is the choice of analytic element used in the analysis – image pixels or 'higher-level' features such as contrast edges or extended contours. These are often termed pixel-based and feature-based processing. At the pixel level, image intensity values are treated in an undifferentiated way, and the resulting representation is often termed "retinotopic" for its resemblance to a retinal layout. Feature-based processing and description works with a distinguished subset of the image information, and leads to scene descriptions that are more sparse but, through better localization, are also more precise. Although in truth this dichotomy is more of a continuum, I will exclusively consider the latter as *structured* abstractions from the imagery – the features will be edge elements or parts of contours.

[1] The question of learning is probably at the root of the question of intelligence.

[2] An immediate question with such analysis lies in what is being tracked through the dynamic sequence, and we will return to a discussion of this.

[3] A variety of choice of 'element' have been developed.

4. SINGLE IMAGE ANALYSIS

A common task in computer vision is to identify or classify items in a single image taken of some scene. For example, the task may be to identify and assemble components of a small machine, or to identify targets in an aerial view of a military installation. Clearly, single snapshot images of such a scene will lack 3D and dynamic information. The processing must rely on some comparison of what the computer expects to see with descriptions it extracts from the single image.

At the pixel level, the comparison may aim to group parts of the scene based on textural and other classifications. For example, a region that exhibits high spatial intensity variation (texture) may be classified as vegetation if the scene is expected to contain vegetation. Homogeneous regions may be sky if, again, the domain is known to be a natural scene out of doors. Anticipated relations between classified regions may provide use of mutual consistency to make the interpretation more robust. For example, if sky must be above vegetation, which is generally above the ground, then these spatial relations should be required of the classified regions. The major determinants of the capability of the system are the quality of the classifiers and the suitability of the relations. One may appreciate that determining effective classifications and relationships, valid across a wide range of realistic situations, might be difficult.

At the feature level, 2D shape descriptors are typically extracted from such imagery, for example straight lines, curves, and smooth contours, grouped into contiguous pieces. Some previous automated or interactive process has led to the development of a 'model vocabulary' – a set of feature groupings that can be composed together to represent the range of objects anticipated in the scene. Recognition involves comparing the extracted features (e.g., lines, arcs) and their interrelationships with those represented by the models.

What is probably most important to observe in this single-image analysis is that the processing must be preceded by defining what is expected to be seen in the images. Since 3D shape and motion are not available to the analysis, recognition must be based solely on the 2D information that can be obtained.

4.1 Interpretation through Pixel Classification

Strat (2) has demonstrated an impressive capability at interpreting natural scenes with a pixel-based classification system along the lines outlined above. He points out that most recognition schemes are based on geometric representations and matching of discrete features, yet natural scenes are neither well described by geometry nor characterized by specific localizable features. Taking a more eclectic approach, he develops a battery of filters that attempt to classify image regions, and builds a relational network among these descriptors. What brings the classifiers together is 'context' – the expected relationships between labeled components. These contexts are established manually in advance of any processing, and are individually constructed for specific domains.

By making the recognition context sets very specific, for example identifying 'foliage against sky' rather than simply 'foliage,' they can be made more reliable. At the same time, generic contexts can be defined that may be satisfied when more specific ones cannot. Context sets may include components that are both positive (for example, tree trunks tend to be vertical), and negative (ground cannot extend above the skyline). A variety of grouping and segmentation techniques are used over a variety of scales to produce candidate scene region labelings – estimates of pixel groupings (similar intensity or color), similar texture, horizontal or vertical orientation, line-like structure, and so forth. Robust operation is attained through use of overlapping or redundant filters. For example, sky may be either an untextured homogeneous region of high intensity or an area of smoothly varying general brightness above most other areas in the image. Cliques – mutually consistent sets of classifications – are sought over the image. The clique providing the greatest reliability and coverage is chosen as the best interpretation of the scene.

Using an auxiliary knowledge representation system (the Core Knowledge System, CKS (3)), a sequence of images may be processed, accumulating and sharing constraints from their individual interpretations. This, together with a coarse use of stereo (4), enables Strat's system to build up a rough symbolic 3D map of the area being viewed.

The examples Strat presents are in outdoors scenes of trees, rolling hills, and pathways. Figure 1 shows a 3D reconstruction of an outdoor scene analyzed with this system.

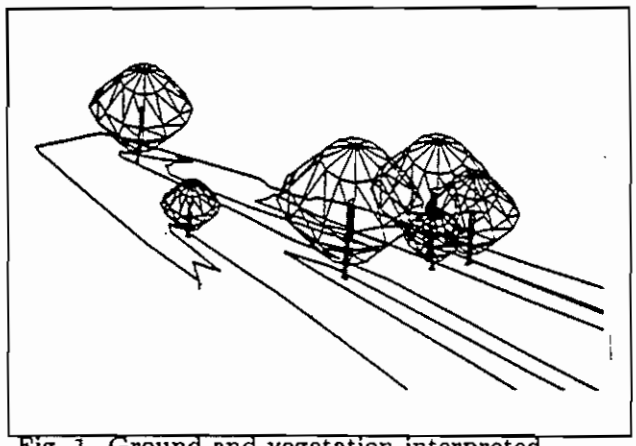

Fig. 1. Ground and vegetation interpreted from a single image.

While demonstrating a good capability at classifying image components in domains where the relationships have been prespecified, this approach is unlikely to provide the depth of interpretation needed for general scene understanding. One factor in this is that the system would require a significantly larger vocabulary of objects with increasingly tight constraints on their interpretation to distinguish, for example, among different types of trees or, more critically, to recognize specific trees, such as the one with a broken branch on the top of a certain hill. This requires geometric understanding rather than an understanding of certain relationships. In addition, no mechanism is presented for abstracting the required rules from the data. If one wants the system to show a utility beyond simple domains, this generative aspect is essential, and geometry probably cannot be avoided. Nevertheless, relational measures are generally missing from geometric-based recognition systems, and the use of this relational

approach in a partnership with the more metric approach of shape- and structure-based techniques should lead to more reliable operation for both.

4.2 Shape from a Single Image

A difficulty in trying to obtain information about shape or 3D structure from a single image is that a particular single image could arise from an infinity of scene configurations. The simplest example of this is an image of the image itself, where there is clearly no three-dimensionality to be observed, only interpreted. Interpretation requires knowledge, including knowledge of the physics of the imaging process and the local implications of intensity variation with respect to the shape of the imaged surface. Nevertheless, we all have the ability to interpret single images as 3D scenes, and there has been considerable effort in the field to develop similar capabilities in the computer. Using iterative optimization techniques and models of illumination, reflectance, and variations including albedo, Leclerc and Bobick (5), and others, have demonstrated the ability to recover surface height from simple measures on the imagery.

That such analysis cannot be guaranteed correct is apparent from its fundamental assumptions. The interplay of reflectances and shadowing could cause havoc with the modeling, which presumes fairly simple relationships between light source and reflecting surface. Any variation is interpreted as either surface shape or simple albedo change. Such shading analysis probably will have its greatest use where other depth measurement techniques, such as binocular stereo, have insufficient information to operate, yet can provide 3D constraint to limit ambiguity.

4.3 Models in Interpreting Single Images

Undoubtedly, much of the world is quite well described geometrically or by discriminable aspects of coloring, texture, or structure. Since the world is three-dimensional, a critical element of scene analysis must be the ability to represent and recognize 3D objects. In these cases, recognition may be attained by locating specific scene features and comparing their parameters with those chosen in advance to represent specific objects. Recognition, here, may be viewed as searching through a set of 3D object descriptions and finding the mapping of position, orientation, and scale that provides the most satisfactory correspondence. Aside from the selection of feature descriptors and the inevitable question of how to acquire the object descriptions in the first place, the major challenge in this work is effective search through the potentially enormous set of match possibilities.

Two pieces of research can highlight the approaches taken to this shape-based or structural recognition. While addressing 3D recognition, each uses information from single images for its recognition. The first represents objects as integrated networks of 3D points. The second provides coverage of the 3D situation by storing a range of representations, each pertaining to a small set of viewing perspectives.

4.3.1 3D Models with Image Matching in 2D

Huttenlocher and Ullman (6) introduced the term 'alignment' – a method to match stored models with features obtained from a view of a scene. In their work, the features – both in the scene and in the model – are two-dimensional contours (each classified by its shape) and their endpoints, if a straight contour, or midpoints otherwise. A model is a set of 3D points forming triangles (planar facets), and the contours of which they are part. Alignment is the process of selecting pairs of corresponding triangles (from the model base and from the imagery) and using the transformation implied by their match to map the rest of the contour description. The transformations are simple translations, rotations, and scalings. Estimating the goodness of fit of the resulting transforms enables selection of a 'best' interpretation.

4.3.2 2D Models and Image Matching

Chen and Mulgaonkar (7) address the problem of model-matching using 2D image data in a more methodical and practical manner. While using a related approach to the matching – hypothesizing 'alignment' transforms and mapping the related constraints for validation with the data, the detail of their strategy offers considerable advantage.

Two characteristics of their work stand out. First, they build their models in a semiautomated way by showing the system parts from various perspectives and under different lighting conditions. Model acquisition is a crucial and potentially[4] very time-consuming component of setting up a recognition task, and a which technique that automates this using the results of its own analysis immediately has more utility. Each model is structured as a set of classified contour elements – straight and curved segments – ordered by their relevance to the matching task. Features that are detectable most often in the training set and are found most likely to be correctly identified in the data are ranked higher in importance. These should be the first to be sought in the matching. This 'learning' strategy enables each model to be organized in a manner that is most effective for establishing its presence or absence in the scene. In effect, a model is a sequence of instructions for validating an object's presence in the image – it is a program.

Their representational system is 2D, and a single object will be composed of several perspective models, with each covering a small range of viewing angles – plus or minus perhaps 15 degrees in each direction. This is not as satisfying a solution as building a unified 3D model of each object; however, it has practical advantages in that it simplifies both the modeling task and recognition.

The system was developed and demonstrated on an industrial assembly operation, involving about two dozen parts, and has since been used for identifying objects in a dynamic context (see subsection 6.3.3).

4.4 Prospect Beyond Single Images

The techniques described above have relied primarily, if not totally, on 2D information, both in their models and in their image understanding. The use of 3D information for model representation and recognition has had less and generally more recent investigation. The principal difference in these works arises from the necessity of obtaining 3D information from the scene. This cannot be done from

[4] "potentially" because very few object recognition system have any sizeable model repertoire

single images, and requires either active ranging (for example, structured lighting, sonar, radar) or at least two simultaneous perspectives from passive sensors such as cameras.

This step to three dimensions lays the foundation for the distinction I wish to make in approaches to image understanding. If the system has no recourse to 3D temporal or spatial information, then its knowledge is limited to what the developer programs in: if the system has an ability to integrate information across space or time, then it can begin to meaningfully augment its knowledge base. Acquisition of this 3D information is the focus of the next two sections.

5. SCENE MODELING FROM STEREO

Image pairs, providing two perspectives of a scene, provide the data for inferring the range to points in a scene. This is termed binocular 'stereo' processing, after its resulting solid three-space description of the scene. The goal of stereo analysis is to obtain the best estimate possible of the range to points in the scene. 'Best' may depend on a number of requirements, including speed. The point to observe about these systems, however, is that they have some knowledge about the state of the world they are looking at – knowledge that serves to constrain the solution they present – and they have the common goal of developing a 3D description of the scene. It is common in stereo research to produce a range map, but very uncommon to do anything further with it, for example, navigating or controlling a robot arm.

Once the camera position and correspondences are known, estimating the range to some feature in the scene is a simple matter of triangulation. An effective mechanism for limiting the cost of determining these correspondences lies in using the 'epipolar constraint.' Knowing the two camera relative positions and attitudes enables definition of the expected pattern of disparity on the images. For cameras directed in parallel, the disparities will only be lateral, while for converging cameras the patterns will be radial. This camera information is used to shape the search window for possible corresponding elements, so it both reduces ambiguity and decreases computational cost.

5.1 Pixels versus Features

Within stereo processing, two major approaches are taken in selecting correspondences, one based at the pixel level and the other at the feature level. The objective within the two is the same, however – recovering the 3D structure of the scene as represented by the 3D location of its components. The main distinction lies in what constitutes these 'components.'

5.2 Scene Geometry from Imnge Pair Pixels

In pixel-based stereo processing, the objective is to label each point in an image (where possible) with a range value. If the relative positions of the cameras are known and corresponding pixels can be found in the two views, then relative range can be estimated directly by triangulation. Absolute range comes from knowing absolute camera displacements. The techniques used for solving the correspondence problem generally involve correlation – estimating the similarity between image regions in the two views. This similarity is usually measured as a local difference in intensity value between corresponding parts of the two images, with secondary constraints being introduced to enforce global consistency. The former, local measure, uses a small support function – typically a square or circular region centered on a pixel – with the similarity being either a simple sum-of-squared differences (SSD), or a correlation coefficient measure. The correlation coefficient measure may be normalized to eliminate the effect of linear variations that might arise, for example, from viewing at different times of the day, under differing light conditions, or with separate automatic gain adjustments on the two cameras.

In SSD matching, the expression to be minimized at any pixel (x, y) is:

$$SSD_{x,y} = \sum_{r_x, r_y} [I_L(x+r_x, y+r_y) - I_R(x+d_x+r_x, y+d_y+r_y)]^2$$

where (d_x, d_y) is a displacement from the source image pixel $I_L(x, y)$, and (r_x, r_y) defines a region of integration in the destination image, $I_R(x + d_x, y + d_y)$. This sum may be weighted to diminish the effect of brightness variance with radius. The vector (d_x, d_y) with minimal sum $SSD_{x,y}$ is selected as the image of the pixel at (x, y) in the second frame.

In normalized correlation, optimization is based on the measure:

$$E = \frac{\sum_{r_x, r_y} [I_L(x, y) - \hat{I}_L][I_R(x, y) - \hat{I}_R]}{\sqrt{\sum_{r_x, r_y} [I_L(x, y) - \hat{I}_L]^2 \sum_{r_x, r_y} [I_R(x, y) - \hat{I}_R]^2}}$$

where $\hat{I}$ is the mean brightness over the image region (r_x, r_y) centered at (x, y).

5.2.1 Normalized Cross Correlation

A typical approach to pixel-based stereo analysis is that of Hannah(4). Here, normalized correlation provides the matching metric, and processing in a resolution hierarchy provides a global consistency constraint. This use of a resolution hierarchy is fairly common in computer vision. It involves building a pyramid-like structuring of the image data, with the bottom level being the full-dimensioned image, and successively higher levels being the half-resolution versions of the one below them. The top level is a small, very highly reduced, and subsampled version of the original image – it has only very low spatial frequency components, with the higher frequencies being removed by the successive averagings.

A strategy often used in computer stereo vision is to match coarse features first (low spatial frequencies), and then use the results at this scale to constrain finer scale matching (higher spatial frequencies).[5] Beyond this constraint, Hannah also requires that her correspondences are the same in left-to-right matches as they are in right-to-left matches. Analysis of the correlation coefficient and

[5] It is always possible to show images in which such an arbitrary direction of progression will give the wrong answer.

an autocorrelation measure enables this process to ignore matches that have insufficient evidence for reliable estimation. This has the benefit that hallucinations, such as giving range to the sky, do not occur often. This technique, however, is costly in computation.

5.2.2 Stochastic Stereo

An alternate that is particularly suitable for implementation on a SIMD parallel processor is a stochastic method, developed by Barnard, using a simulation of the physical process of annealing to enforce global consistency (8). This method uses a composite similarity measure – image intensity difference and a gradient constraint that biases the solution in favor of a flat disparity map. The stochastic element enters the analysis in the way the individual difference measures are combined in looking for a global solution for the image pair. As in annealing, the system is injected with energy (heat), allowed to cool, heated up again – although less – then cooled again, repeating until there is very little change between these heat/cool cycles. The measured change is this similarity measure – a weighted sum of intensity difference and implied disparity gradient for the selected pixel matches. The different 'heat' settings allow a varying range of disparity adjustments in the pixel matching.

The measure minimized for optimization in stochastic stereo is:

$$E_{ij} = \sum_{ij} (|\Delta I_{ij}| + \lambda |\nabla D_{ij}|),$$

with $\Delta I_{ij} = I_R(i, j + D_{ij})$, where I_L and I_R are the left and right brightness values, and ∇D_{ij} is the gradient of the associated disparity estimate; λ balances the brightness and smoothness constraints.

Even when a parallel processor is used, the cost of iteration makes this a fairly time-consuming technique. Images of size 512 by 512 pixels require about 10 minutes of processing time on an 8000-processor Connection Machine (CM).

5.2.3 Real-Time SSD Matching

A third technique worth examining for its simplicity and effectiveness is an SSD method implemented on both a 16000-processor CM and on a coarse-grained (5-processor) i860 parallel processing system (9). Much effort was invested in making this process run as rapidly as possible to support real-time control, and it can perform stereo matching on images 256 pixels square at about 40 Hz on the CM and 10 Hz on the i860 configuration. The SSD phase gives velocity estimates for each pixel, mode analysis of this velocity distribution selects the major discrete motions, and an adjustment phase tracks regions over time. It has been used to control a robotic arm in tasks such as maintaining centered view on pedestrians and on another robot arm.

5.2.4 Considerations

Both of these parallel approaches share a common drawback. They process only in integer units of disparity, so deliver just a small number of bits of range resolution. In the case of the stochastic stereo, this was about 5 bits (32 levels), while with the SSD method it was about 3 bits (8 levels). Any change in this precision incurs added computational cost. Hannah's method delivered subpixel correlation measures, and was precise down to small fractions of a pixel unit.

5.3 Structured Stereo Processing

Another approach to stereo analysis for obtaining 3D information about a scene involves the processing of not pixel values but abstracted features – contour elements as produced by zero-crossing operators. Marr and Poggio, Baker, and Mayhew and Frisby were the early developers of this feature-based approach to stereo matching.

Marr and Poggio (10), later joined by Grimson (11), worked with zero crossings of the Laplacian of a Gaussian (LOG), and progressed from large Gaussians to small Gaussians in a hierarchic-pyramid manner. Matches obtained at the coarse level constrained the possible matches at finer levels. A consistency measure was implemented by insisting that disparities over a small region were identical. An unfortunate artifact of this is that their results tend to represent the scene as planar chunks at different ranges. Mayhew and Frisby (12), later joined by Pollard (13), used a figural continuity constraint to enforce connectivity of depth estimates for LOG features that were connected in projection. They also used peaks and troughs of this signal, presenting evidence from psychophysics supporting human use of these in vision, and introduced a variation of the scale analysis of Marr and Poggio – looking for consensus in neighboring bands rather than in successive coarse-to-fine levels. Baker (14) used a form of figural continuity as well, and followed his feature matching (extrema of intensity gradient related to zeros of the LOG) with constrained intensity matching to provide a dense range map. Grimson used a surface-fitting technique to interpolate between matched features to estimate this map.

The fact that feature-based stereo results in sparse range measures has been raised as a criticism. Dense results are preferred. Feature-based approaches have greater precision, however, as they focus on the more localizable parts of the imagery. Scale processing is felt to be a key to providing dense results. Pixel-based techniques have been more easy to implement on SIMD parallel processors, so they may have an inherent advantage for real-time development.

Much other research has addressed pixel-based and feature-based stereo, including using a third camera to provide an ambiguity-resolving perspective and introducing other constraints (a recent survey paper covers much of this area well (15)). Among some dozen and a half systems evaluated competitively a few years ago (16), Hannah's system was ranked first across a majority of the categories (17).

5.4 Differential Techniques: Motion and Range

A different approach to disparity estimation has been developed for motion processing – optic-flow analysis – where the objective is to estimate movements in a scene (18). Under certain conditions these techniques may also be used for stereo range estimation. Two principal points distinguish this work from pixel- and feature-based

matching approaches. First, the presumption is that there is very little difference from one image to the next – motion processing allows this, whereas typical stereo has a sufficiently large baseline that images may differ significantly. Second, differential techniques are used that do not depend on feature localization in the image.

5.4.1 Optic-Flow Analysis

Horn and Schunk (19) developed the brightness-constancy constraint, which relates variation of intensity between successive images with the underlying variation in the scene. The principle behind this differential technique is that derivatives of the spatiotemporal intensity data indicate rate of image change. If the image change is due only to camera displacement, then simple derivative convolutions on the spatiotemporal intensity data can be used to estimate scene distances. If the change is due to scene motion, then the technique estimates velocities. Since the expression for the variation at a single point is underconstrained, the solution involves a least-squares approximation that integrates over some local neighborhood, and this makes the result sensitive to the density of discrete motions in the vicinity. The estimates are best where there is strong local texture (surface detail) with a single velocity. Where the texture is weak (there is little distinctive detail) or the local vicinity contains more than one motion (such as occurs at object boundaries), the estimate can be rather meaningless. Despite this, the results tend to be generally credible.

With the differential approach, image disparity (or velocity) (d_x, d_y) at frame t can be determined by minimizing the following expression:

$$\sum_{r_x, r_y} [d_x I'_x(x,y,t) + d_y I'_y(x,y,t) + I'_t(x,y,t)]^2,$$

where I'_x, I'_y, and I'_t are spatial and temporal derivatives of image intensity $I(x,y,t)$.

The summation is again taken over a local region of the image (r_x, r_y). One finds the least-squares solution, in closed form, by taking derivatives of this expression with respect to d_x and d_y. The least-squares estimate is given by:

$$\hat{\mathbf{d}} = -\mathbf{M}^{-1}\mathbf{b},$$

where

$$\mathbf{M} = \begin{pmatrix} \sum I'^2_x & \sum I'_x I'_y \\ \sum I'_x I'_y & \sum I'^2_y \end{pmatrix},$$

and

$$\mathbf{b} = \begin{pmatrix} \sum I'_x I'_t \\ \sum I'_y I'_t \end{pmatrix}.$$

This expression has minimum error when

$$d_x I'_x + d_y I'_y + I'_t = 0,$$

that is, when the observed image gradient vector (I'_x, I'_y, I'_t) is orthogonal to the observed disparity (or velocity) vector $(d_x, d_y, 1)$. Figure 2 shows the optic flow computed for the motions of a sedan and van against a stationary background, the imagery of which is shown at the top of Figure 5.

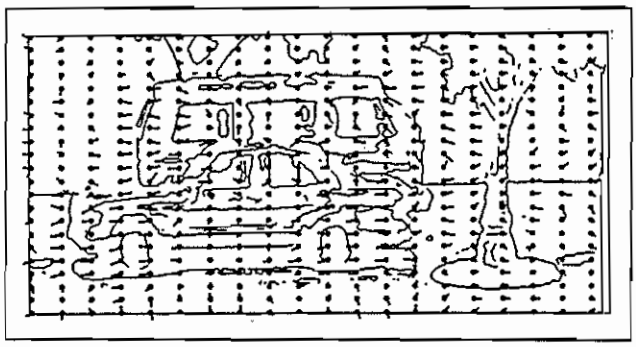

Fig. 2. Optic Flow for Moving Sedan and Van.

5.4.2 Hierarchic Optic-Flow Computation

Hanna has presented a method for extending the applicability of the gradient-based technique to images with significant variation between frames (20). This operates through a hierarchic-pyramid analysis, beginning with low-resolution coarsely sampled imagery, and progressing through to the full resolution data. A unit of pixel measure in the coarse imagery corresponds to a 2^n by 2^n pixel region at highest resolution n ¡levels finer, so a gradient computed at this single unit can identify the predominant motion over that much larger window. Recursive processing of this motion estimation followed by image remapping – to bring the corresponding image locales into alignment for the next gradient analysis – may be viewed as delivering the n-bit motion vector a bit at a time, starting from the highest-order bit. What is important to note is that with this hierarchic approach, gradient-based optic flow can also be used for stereo range estimation – large disparities are handled by the coarser scales. The major difficulty remains, however, that there can be no guarantee this coarse-to-fine progression will give correct results. A small feature that is moving to the left while the predominant region motion at a coarse level moves to the right will be 'mapped' in the wrong direction for being detected at any of the succeeding levels.

An iterative remapping method very similar to Hanna's was used much earlier by Quam in his *hierarchical warp stereo* process (21). The matching metric in this work was correlation, rather than gradient-based optic flow.

5.5 Issues in Stereo Processing

A number of questions must follow any depth recovery process, such as: Are there measures of confidence associated with individual estimates? Is the result conclusive? Are there errors of omission (gaps) or commission (range estimates where there can be none)? Does the process deliver a description of objects or just an array of numbers that represent a range 'map?' How relevant is the resulting description to the intended use? Since the purpose of range recovery is tied to some other task, such as understanding the scene or moving about in it, these questions can determine the utility of the whole exercise.

One of the principal dissatisfactions in stereo analysis has been in its reliability. Perhaps 90% of a scene can be adequately modeled with the above techniques, but the remaining 10% failure can make the results almost unusable. Higher reliability is needed before one can trust an autonomous device for guidance. There is very little opportunity to obtain better accuracy when presented

with only two perspectives of a scene. Ambiguities are difficult to detect, and cannot be resolved without the introduction of more information. This information has often taken the form of *a priori* knowledge about scene and object types (for example, that the scene contains static opaque rectilinear structures).

Better additional information that is not domain specific, is provided by "trinocular stereo," which involves acquiring a third view of the scene. This was first introduced by Burr (22) and later followed by Faugeras's group in France (23). This third view, if noncollinear with the other two, provides a second epipolar constraint that can disambiguate potential match uncertainties.

Almost without exception, stereo techniques have difficulty in correct handling of occlusion (where a feature does not have a match in the corresponding view), image reversals (where feature left-to-right ordering is inverted between views), transparency (where multiple ranges are associated with individual view points), and canopy phenomena (where there are a few predominant and quite different depth ranges over a small region of the view). These are significant issues for depth estimation and natural scene interpretation.

A more general comment on two- or three-view stereo is that the resulting descriptions are not of the same quality as those we perceive when we as humans observe a scene. Stereo results look like cut-outs, with a series of ranges computed for certain directions of the camera. The same can be observed in looking at a stereo pair of photographs – the perception is likely to have a flat, disjoint, and chunky appearance. The perception we have under natural conditions is more continuous and connected, and this results from our ability to observe in the continuum through time. We change our viewing position to suit our demands for fill-in and clarification, and integrate information through active control of the viewing process, such as obtaining a description of some novel 3D object by grasping it and manipulating it before the eyes.

6. SCENE MODELING FROM SEQUENCES

Recent approaches to 3D vision have addressed this processing of image sequences, where a sequence comprises many views from different positions. This more closely resembles the operation of the human system, where we observe with eyes that are free to move, collecting information from various perspectives. This multiple-view approach could provide considerably more complete descriptions of a scene, revealing, for example, what the back side of an object looks like, and could do so with much less ambiguity. Aside from restricted cases, however, it has proved difficult to exploit this extra data in the coherent manner required. One of the problems lies in organizing and maintaining coherent descriptions of the rather massive amount of data involved – sequences could be hundreds of frames long, or more.

6.1 Correspondence Through Time

Sequence processing shares many of the computational issues of stereo. The principal problem in stereo processing has been identified as putting into correspondence, accurately and reliably, features that appear in two views of a scene. Determining the correspondence is an ill-posed problem: ambiguity, occlusion, image noise, and other influences resulting from the differing appearance of objects in the two views make feature matching difficult. In sequence analysis, where rapid image sampling produces images that change little from one to the next, matching is less problematic. In some approaches this is taken to an extreme, with sampling sufficiently rapid that images vary smoothly between views. The following sections describe how this temporal continuity has been developed and exploited for robust tracking and estimation of scene features.

6.2 Pixel-Based Sequence Analysis

As was the case with stereo analysis (cross-correlation and gradient analysis), there are two principal approaches to pixel-based motion analysis. In correlation, the objective is to determine for each pixel in one frame, its image in the next frame. Techniques as described in section 5.2 are used for this. SSD is more typical than normalized correlation in sequence analysis. With temporal sampling sufficiently fine that brightness changes are of a smaller magnitude than changes due to motion, there is little requirement for accommodating to varying illumination. With the optic-flow approach, on the other hand, explicit matching is avoided, and motion is derived directly through differential analysis, as described in section 5.4.

Another problem both correlation and optic-flow analyses encounter is that they are designed for pair-wise computation rather than for sequential tracking. Since they are referenced on the center of a pixel in one image, their displacements are not easily chained with precision through a sequence. Range estimates will be imprecise over a short baseline, so the reliability and precision obtainable for matches over a long baseline become crucial questions.

Pixel-based and point-based reconstruction techniques, where they have been developed to the stage of integrating measures over a sequence (for example, (24, 25)), do not exploit the continuity of observations. Rather, they treat observations from different perspectives as disjoint, and pool them in (more or less estimation-theoretic) volume sets.

A recent innovation – the use of a singular value decomposition procedure – uses intermediate feature trackings to synthesize a long baseline through many small changes. It recovers both the shape and motion observed in transformation of a rigid body (26). The tracking employed uses an autocorrelation measure to select distinctive image features (in a spirit similar to that of Hannah). By tying observations together through the sequence, it obtains the benefits of a large baseline with the reduced error of small-increment image variation.

A difficulty with local-support integration techniques (pixel-based approaches in general) is that when the local region of integration overlaps different range distributions, the estimate may be quite meaningless. Since these bounding areas are of particular interest in most 3D tasks – such as grasping and navigating – this deficiency can be quite severe. The issue is particularly salient in motion analysis, where an intermediate velocity estimate is much

more misleading than an intermediate range estimate. Intelligent window shaping may improve the situation, although at significant cost (27).

6.3 Structured Processing – EPI Analysis

There is much more in an image sequence than is being processed by techniques such as those described above. Selecting only highly localizable features leads to sparse scene descriptions, while use of the full image contents, as in optic-flow and correlation approaches, leads to much uncertainty, weak localization, and fragmented tracking. An alternative exists in utilizing the three-space correlate of 2D image contours. The motivation of this 'structured' approach to sequence analysis is that dynamic imagery has both spatial and temporal structure, while pixel-based techniques represent neither and must determine them both during its operation. Pixel-based techniques compute the temporal structure by 'tracking' features using correlation or optic-flow analysis, and determine the spatial structure by grouping results after temporal tracking. And yet the structure is there in the data.

Epipolar Plane Image (EPI) Analysis is such a technique that holds particular promise for scene reconstruction (28). It integrates throughout the data acquisition and has several major advantages over other approaches, such as not requiring correlation or any similar matching strategy, and dealing explicitly with spatial and temporal continuity. The features utilized are at object and texture discontinuities, so do not involve integration across different range distributions. This technique was the first to exploit small increments over a large integrated continuous baseline for the ideal mix of reliability and precision in motion analysis. The geometry and intuition of imaging in this situation are a little unusual, so I will review the implications of the generally used epipolar constraint in the context of sequence processing.

6.3.1 Epipolar Geometry

In Figure 3 (left), a camera is shown at two different positions along a linear path. At each of the sites the camera is looking at right angles to the path, and a feature such as P will appear displaced to the right in the second view with respect to the first. This displacement is along the projection of the plane formed by P and the two camera centers. This plane is termed an "epipolar plane." For a continuing sequence of such images, the point P will stay on the same image scan line from frame to frame. Because of this epipolar structuring, we can confine our depth analyses in right-angled linear motions to single sets of scan lines. Figure 4 shows a volume formed by stacking up the data collected in an image sequence and slicing horizontally to reveal such a set of scan lines. The pattern of streaks in this slice makes the lateral displacement character quite apparent and their interpretation quite direct: Near features have streaks with low slopes, more distant features have higher slope. Stereo processing of such a scene would correspond to comparing features between, say, the first and the last frame, or the first and last line of this image. The continuity evidenced here takes the uncertainty out of the matching process. Analysis of these slice images, termed epipolar-plane images (EPI images) after their composition from samples of a single epipolar plane, led to an effective technique for estimating the range to features in a scene.

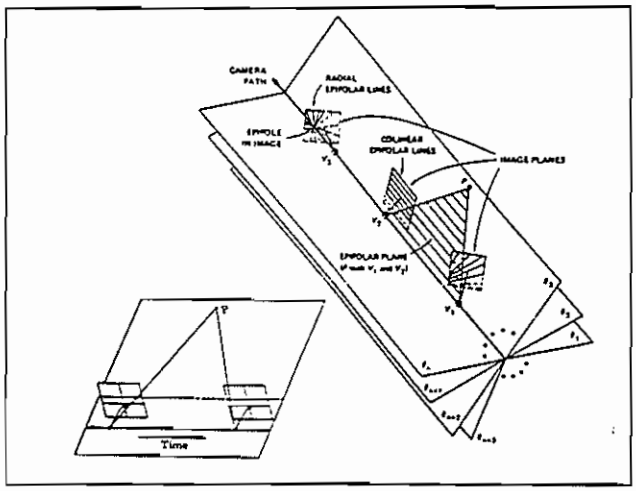

Fig. 3. Epipolar Configuration for Moving Camera.

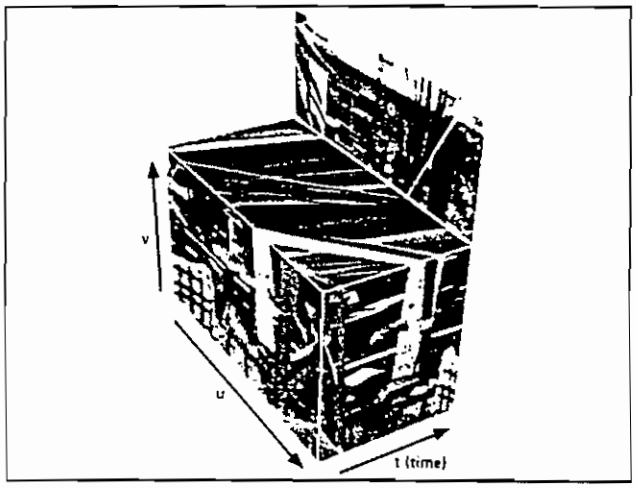

Fig. 4. Spatiotemporal Image Volume.

6.3.2 Spatiotemporal Manifolds

To expand the technique to more complex viewing situations such as nonlinear and varying-velocity camera paths with varying camera orientations, as would be found when a human moves through a scene (Figure 3 (right) shows patterns of epipolar lines that arise for linear motion and varying view direction), it was necessary to generalize the geometric representations used. In the earlier work, EPI-based linear features – representing the evolution of individual features over time – were detected and processed. In generalizing the approach, *spatiotemporal manifolds* – representing the time evolution of whole spatial contours – were constructed and used in inferring scene structure (29).

This reformulation brought another advantage: Representing the time-evolution of contours rather than individual features would produce connected 3D space curves rather than isolated points. Grouping of scene measures into meaningful and related structures remains one the largest problems in vision. Since even the most reliable and precise depth map is only another input to the scene-understanding process, any technique that can deliver direct segmentation and grouping information with its measures will have a great impact on the use and reliability of its data.

6.3.3 Tracking and Identification

Figure 5 shows a composite development in tracking and identification using the spatiotemporal manifolds for feature localization in space and time, and the 2D modeling facility of Chen (7) for object recognition. The figure shows in successive steps the strongest zero-crossing contours in three adjacent frames (the first and last of which are shown at the top), with the final view showing the results of identifying a van and sedan in these data. The bottom of the figure shows the models used in the recognition. These were constructed in a earlier training phase. An added benefit in this figure is that it demonstrates the value of stereo in perception: The paired figures are presented for crossed-eye viewing and, when fused into a single percept, will reveal a considerably more coherent interpretation, one that may be impossible to obtain monocularly.

6.4 Stereo and Motion

Undoubtedly, simultaneous stereo and motion analysis must be obtained for us to hope to achieve the capabilities of the human mobile-binocular system. Stereo is essential, as motion can only compute range to stationary objects and for known camera motion. At the same time, motion and sequence analysis are essential, as the active element in exploring an environment, both for modeling it and for navigating through it, cannot be met from a single perspective or even a set of predetermined perspectives. While the number of research efforts addressing stereo and motion analysis is small (9, 24, 25, 30), a coherent approach to integrating these two related modalities will be essential to capturing the true three-dimensionality of our environment. Figure 6 shows an integration of this sort of stereo range estimation and sequence processing operating on a field of rocks. The initial description (middle) is refined from subsequent views resulting in better definition on object 3D shape (bottom). The computational requirements for this data-intensive challenge are now being met by multi- and parallel-processors, with a number of research groups investigating stereo sequence analysis in high-performance computing environments.

6.5 Recognition of 3D Shape

The techniques described above have addressed the issue of obtaining estimates of scene 3D structure from two or more views. The major purpose of this is to provide the third dimension for tasks involving recognition and navigation. Unfortunately, very little has been done in using the 3D estimates produced. An early effort that took on this problem was my modeling research in Edinburgh (31). Models of 3D shape were constructed through analysis of objects observed rotating about a known axis. Using a 3D alignment technique, models built from current imagery were compared with models stored in the training phase, and the closest 3D fit was selected as the match.

Although more refined techniques have been developed in the interim, for example the work of Szeliski (32) in building 3D representations using rotation, the majority of research in 3D model matching has used either very simple representations, such as rectilinear blocks (33), or direct ranging techniques, such as provided by structured light or laser devices (34). Where 3D objects have been recognized, they have rarely been modeled by the same process used for their recognition. An exception to this lack of acquisition and use of 3D information in computer vision is in autonomous navigation systems (35, 36), although most systems use active ranging. Some of these systems are capable of extracting 3D scene features and then using these in obstacle-avoiding traversal of the area. Again, however, the representations tend to be simple (boxes, points) and not adequate for representing anything of the sophistication and detail of our environments. A good review of 3D object description techniques may be found in a paper by Besl (37). Some of the works he cites address the issue of model building within a recognition context.

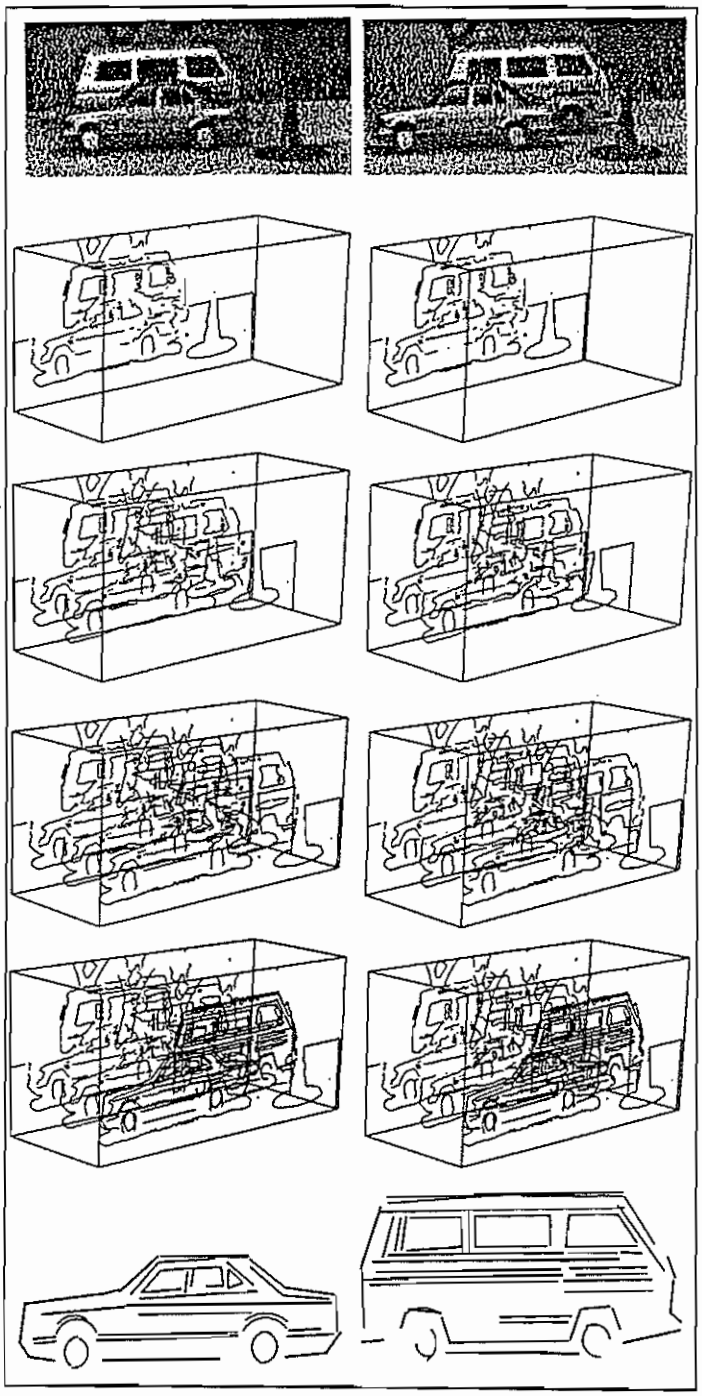

Fig. 5. Object Recognition in Spatiotemporal Tracking.

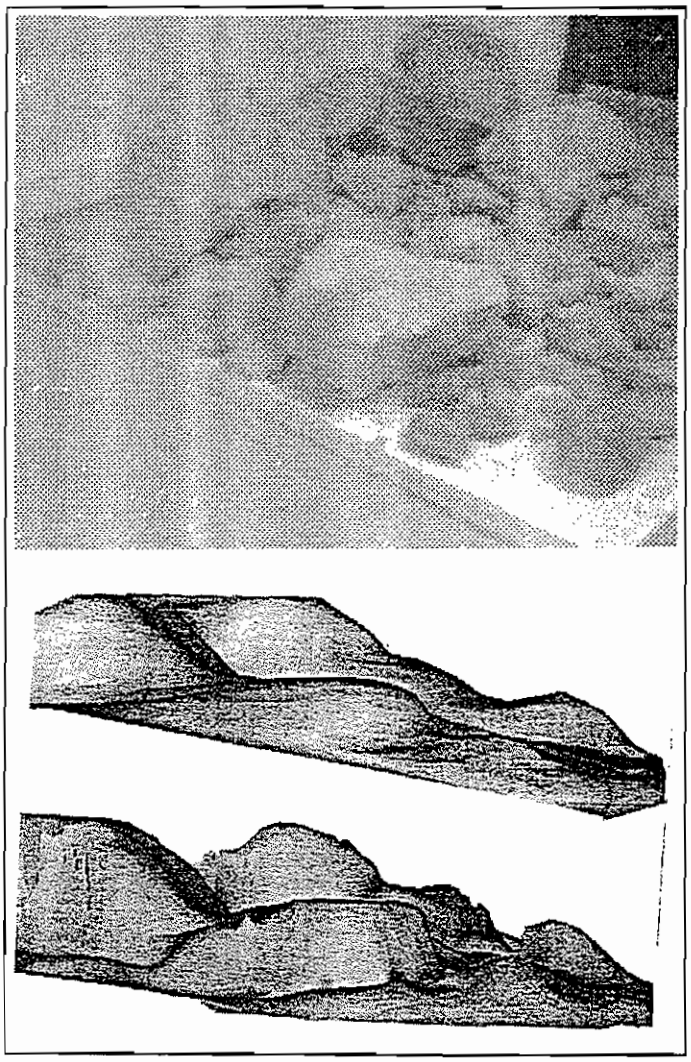

Fig. 6. Refined Scene Model from Stereo Sequence.

7. CONCLUDING REMARKS

A system that is to operate in the real world – that is, to find its way around and interact with other processes in the environment – must be able both to use information about the scene and to derive information during its operations through use of its sensors. This building and using of information in scene analysis, both geometric and otherwise, is an essential element for autonomous operation. Given sufficiently expressive modeling, single images will be adequate for interpretation, but to capture these models requires developing temporal and stereo integration techniques, and ones that encompass both geometric and relational information about objects and their surroundings. The alternative – programming in advance whatever is to be seen – cannot deliver the flexible capabilities needed for operation in the relatively unstructured and unconstrained domains in which we hope to operate our vision systems.

When looking at the challenge of precision operation in a world with the complexity of ours, we can see we have come a long way, yet still have considerably more to accomplish. Techniques for analysis over scale, 2D and 3D object modeling, optic-flow and spatiotemporal analyses, combining with object recognition using 2D and 3D geometric and relational descriptors, are leading us in the direction of attaining these capabilities.

References

[1] Roberts, L. G. (1965). "Machine Perception of Three-Dimensional Solids," *Optical and Electro-Optical Information Processing*, MIT Press.

[2] Strat, T. M., and M. A. Fischler (1991). "Context-Based Vision: Recognizing Objects Using Information from Both 2-D and 3-D Imagery," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 13, No.10, 1050–1065.

[3] Smith, G., and T. M. Strat (1987). "Information Management in a Sensor-Based Autonomous System." *Proc. DARPA Image Understanding Workshop*, 170-177.

[4] Hannah, M. J. (1980). "Bootstrap Stereo," *Proc. DARPA Image Understanding Workshop*, 201–208.

[5] Leclerc, T. G., and A. F. Bobick (1991). "The Direct Computation of Height from Shading," *Proc. Computer Vision and Pattern Recognition*, Maui, Hawaii, 552–558.

[6] Huttenlocher, D. P., and S. Ullman (1988). "Object Recognition Using Alignment," *Proc. DARPA Image Understanding Workshop*, 370–379.

[7] Chen, C-H., and P. G. Mulgaonkar (1992). "Automatic Vision Programming," *Computer Vision, Graphics and Image Processing: Image Understanding*, Vol. 55, No.2, 170–183.

[8] Barnard, S. (1989). "Stochastic Stereo Matching Over Scale," *Intl. Jour. Computer Vision*, Vol. 1:1, 17–32.

[9] Woodfill, J. L., and R. D. Zabih (1991). "An Algorithm for Real-time Tracking of Non-Rigid Objects," *Proc. American Assoc. Artificial Intelligence*, Anaheim, CA., 718–723.

[10] Marr, D., and T. Poggio (1979). "A Computational Theory of Human Stereo Vision," *Proc. Royal Society of London*, Vol. B204, 301–328.

[11] Grimson, W. E. L. (1981). "A Computer Implementation of a Theory of Human Stereo Vision". *Proc. Royal Society of London*, Vol. B292, 217–253.

[12] Mayhew, J. E. W., and J. P. Frisby (1981). "Psychological and Computational Studies Towards a Theory of Human Stereopsis," *Artificial Intelligence*, Vol. 17, 349–385.

[13] Pollard, S. B., J. E. W. Mayhew, and J. P. Frisby, (1981). "PMF: A Stereo Correspondence Algorithm Using a Disparity Gradient Limit," *Perception*, Vol. 14, 449–470.

[14] Baker, H. H., and T. O. Binford (1981). "Depth from Edge and Intensity Based Stereo," *Proc. Seventh Intl. Joint Conf. Artificial Intelligence*, Vancouver, B.C., 631-636.

[15] Dhond, U. R., and J. K. Aggarwal (1989). "Structure from Stereo – A Review," *IEEE Trans. Systems, Man, and Cybernetics*, Vol. 19, No.6., 1489–1510.

[16] Gülch, E. (1988). "Results of Test on Image Matching of ISPRS Working Group III/4," *Intl. Archives of Photogrammetry and Remote Sensing*, Vol. 27, III, 254–271.

[17] Hannah, M. J. (1988). "Digital Stereo Matching Techniques," *Intl. Archives of Photogrammetry and Remote Sensing*, Vol. 27, III, 280–293.

[18] Heeger, D. J. (1988). "Optical Flow Using Spatiotemporal Filters," *Intl. Jour. Computer Vision*, Vol. 1:4, 279–302.

[19] Horn, B. K. P., and B. G. Schunk (1981). "Determining Optical Flow," *Artificial Intelligence*, Vol. 17, 185–203.

[20] Hanna, K. J. (1991). "Direct Multi-Resolution Estimation of Ego-Motion and Structure from Motion," *IEEE Workshop on Visual Motion*, New Jersey, 156–162.

[21] Quam, L. H. (1983). "Hierarchical Warp Stereo," *Proc. DARPA Image Understanding Workshop*, 149–155.

[22] Burr, D. J., and R. T. Chien (1977). "A System for Stereo Computer Vision with Geometric Models," *Proc. Fifth Intl. Joint Conf. Artificial Intelligence*, Cambridge, Mass., 583.

[23] Ayache, N., and F. Lustman (1987). "Fast and Reliable Passive Trinocular Stereovision," *Proc. Intl. Conf. Computer Vision*, London, 422–427.

[24] Grosso, E., G. Sandini, and M. Tistarelli (1989). "3-D Object Reconstruction Using Stereo and Motion," *IEEE Trans. Systems, Man, and Cybernetics*, Vol. 19, No.6., 1465–1477.

[25] Fua, P., and P. Sander (1992). "Reconstructing Surfaces from Unstructured 3D Points," *Proc. DARPA Image Understanding Workshop*, San Diego, 615–625.

[26] Tomasi, C., and T. Kanade (1991). "Factoring Image Sequences into Shape and Motion," *IEEE Workshop on Visual Motion*, New Jersey, 21–28.

[27] Okutomi, M., and T. Kanade (1992). "A Locally Adaptive Window for Signal Matching," *Intl. Jour. Computer Vision*, Vol. 7:2, 143–162.

[28] Bolles, R. C., H. H. Baker, and D. H. Marimont (1987). "Epipolar-Plane Image Analysis: An Approach to Determining Structure from Motion," *Intl. Jour. Computer Vision*, Vol. 1:1, 7–55.

[29] Baker, H. H., and R. C. Bolles (1989). "Generalizing Epipolar-Plane Image Analysis on the Spatiotemporal Surface," *Intl. Jour. Computer Vision*, Vol. 3:1, 33–50.

[30] Zhang, Z., O. D. Faugeras (1992). "Three-Dimensional Motion Computation and Object Segmentation in a Long Sequence of Stereo Frames," *Intl. Jour. Computer Vision*, Vol. 7:3, 211–241.

[31] Baker, H. H. (1976). "Three-Dimensional Modelling," *Proc. Fifth Intl. Joint Conf. Artificial Intelligence*, Cambridge, Mass., 649–655.

[32] Szeliski, R. (1990). "Shape from Rotation," *Proc. Computer Vision and Pattern Recognition*, Maui, Hawaii, 625–630.

[33] Lowe, D. L. (1990). "Integrated Treatment of Matching and Measurement Errors for Robust Model-Based Motion Tracking," *Proc. Intl. Conf. Computer Vision*, Osaka, 436–440.

[34] Chen, C-H., and A. C. Kak (1989). "A Robot Vision System for Recognizing 3-D Objects in Low-order Polynomial Time," *IEEE Trans. System, Man, and Cybernetics* Vol. 19, No.6, 1535–1563.

[35] Ayache, N., and O. D. Faugeras (1989). "Maintaining Representations of the Environment of a Mobile Robot," *IEEE Trans. Robotics and Automation,*, Vol. 5, No.6, 804–819.

[36] Iyengar, S. S., and A. Elfes, editors (1991). Autonomous Mobile Robots: Perception, Mapping, and Navigation, IEEE Press, Washington.

[37] Besl, P. J., and R. C. Jain (1985). "Three-Dimensional Object Recognition." *Computing Surveys*, Vol. 17, No.1, 75–145.